# 03 | 技术视角：AI产品经理需要懂的技术全景图

2020-12-18 刘海丰

成为AI产品经理 进入课程 >

**讲述：刘海丰**
时长 18:46 大小 17.22M

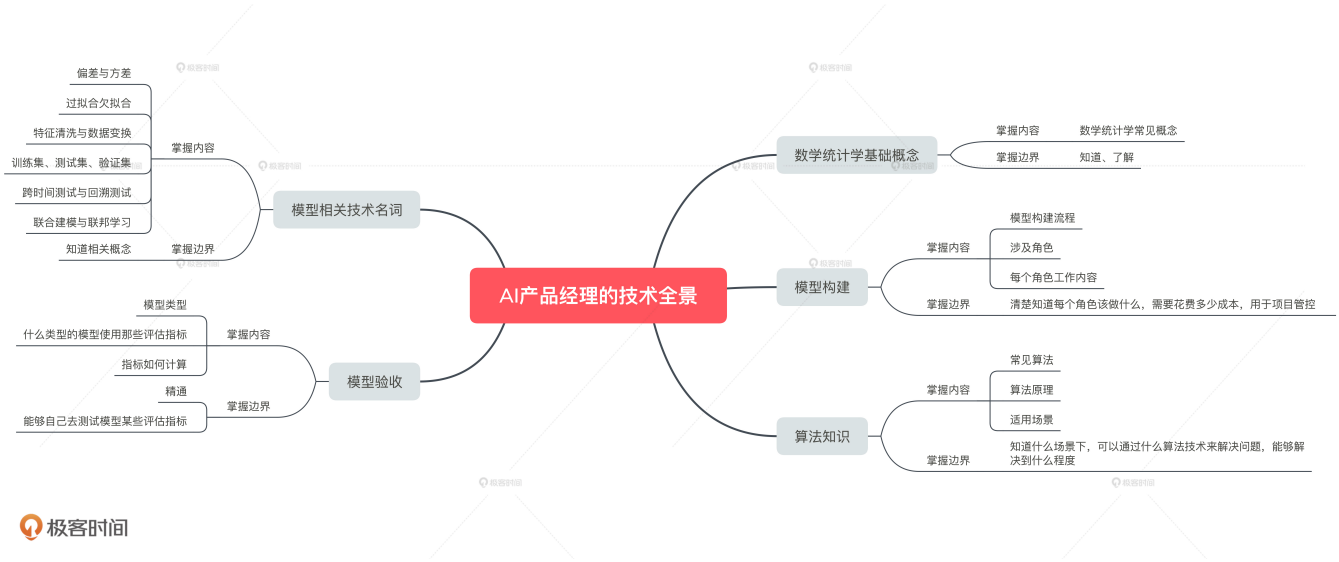
你好，我是海丰。今天，我想和你聊聊 AI 产品经理需要懂的技术有哪些。

在转型成为 AI 产品经理之前，你一定最关心技术问题。这也是很多同学的疑问，比如：AI 产品经理要不要懂技术？现在市面上 AI 的课程特别多，但都面向算法工程师，内容太复杂，我到底要掌握到哪种程度呢？

这些疑问我现在就能给你一个肯定的回答：AI 产品经理一定要懂技术。这就像一个产品经理要懂研发技术是一样的道理。但是，我们具体要掌握哪些技术，掌握到什么程度，是不是要像技术人员一样去学习市面上那些 AI 课程呢？这就是我们今天要解决的问题。

这节课，我会站在 AI 产品经理的视角，结合一张 AI 技术全景图，来帮你解决学什么技术和学到什么程度的问题。让你能够和算法工程师同频沟通，知道该如何去管控 AI 项目进

度，让你最终有能力去牵头主导一个 AI 项目。




总的来说，AI 产品经理需要知道五方面的技术知识，分别是数学统计学相关的基本概念，模型构建的整个流程，常见算法的原理和应用场景，模型验收的具体指标和方法，以及模型相关的技术名词。其中，模型的构建流程、算法的技术知识和模型的验收标准这三项知识非常重要，它们也是 AI 产品经理必备的核心能力，所以我还会在后面的课程中单独来讲。

## 数学统计学基础概念

首先，我们来看产品经理需要学习的数学统计学知识。你可能会奇怪，作为一个产品经理，为什么还要学数学呢？

因为今天的各种人工智能技术都是建立在数学模型基础之上的，必备的数学统计学知识是理解人工智能的基础，所以作为 AI 产品经理来说，这些基础知识也是必须要学习的。

只不过，**对于 AI 产品经理来说，虽然不需要了解数学公式，以及公式背后的逻辑，但我们需要知道数学统计学的基本概念，以及概念的落地应用**。所以，我不会给你讲导数和偏导数的公式，也不会讲贝叶斯推导，我只会把我在工作中接触到的，比较多的数学和统计学概念整理出来。了解这些概念，知道它们的作用，对我们来说就足够了。

这些数学和统计学的知识可以分成两大类：一类是线性代数中的基础名词，如标量、向量、张量；另一类是概率统计中的常见分布，如正态分布、伯努利分布。接下来，我们先从线性代数中的基础名词讲起。

## 线性代数

我们拿起笔，先在本子上画上一个个小圆点，每个这样的点可以代表一个整数、实数或者复数。这样一个单独的数，在线性代数中我们叫它**标量**（Scalar）。

如果我们把这些标量按一定顺序组成一个序列数，如 {x1 , x2 , x3 ,..., xn}，这样的数列就叫做**向量**（Vector）。你也可以理解成，我们给一个个单独的数（标量）增加了一个维度，它就变成了一个数组（向量），所以向量可以看做标量的扩展。

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

又因为每个向量都由若干标量构成，如果我们把向量中所有标量都替换成相同规格的向量，就会得到一个**矩阵**（Matrix）。同样的，矩阵也可以看作是向量的扩展，是我们给原始的向量增加了一个维度，让它变成了一个二维数组。

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix}$$

到这你是不是有点迷糊了？这样一来，矩阵和标量又有什么关系呢？

比如说，我们可以把矩阵看成一个灰度图像，如果一张灰度图像是由 32*32 个像素点组成，那我们就可以把这个图像看成是一个32*32 的矩阵，里面的每一个像素点就是由灰度值（0 到 255）组成的标量。

再比如说，在我们做用户画像的时候，如果有 N 个用户，每个用户有 M 个特征，利用它们，我们就会得到一个用户画像。它可以看成是一个 N*M 的矩阵，矩阵中的每一个点，都是某一个用户对应的某个特征，是一个具体的数值，也就是标量了。

如果我们将矩阵中的每个标量元素再替换成为向量的话，就会得到一个**张量**（Tensor）。这个时候，张量就可以看作是矩阵的扩展，是给原始的矩阵增加了一个或多个维度之后得到的。我们也可以把张量理解成是矩阵向任意维度的扩展，它是深度学习框架中的基本概念。

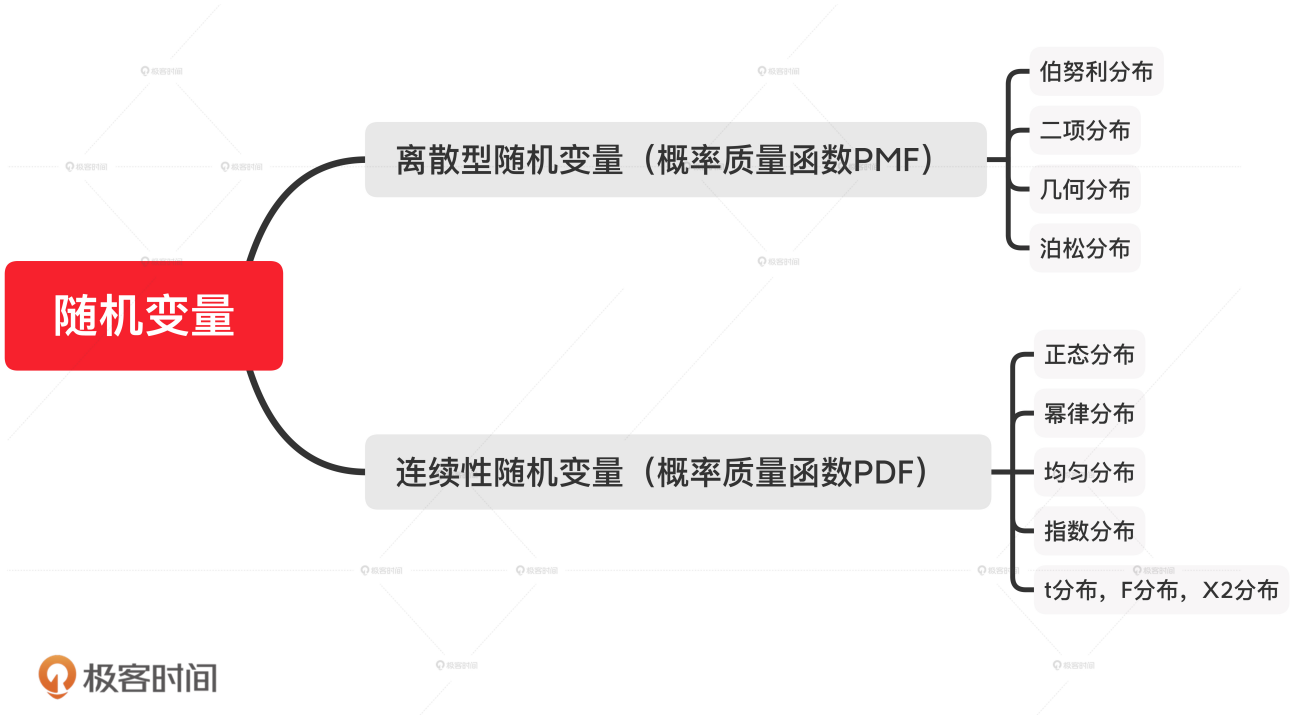
到这里，你可能想问，概念这么多，怎么才能记得住啊？这里我再教你一个小技巧。我们可以把标量看成是零阶张量，向量看成是一阶张量（一维数组），矩阵看成是二阶张量（二维数组），而且任意一张彩色图片都可以表示成一个三阶张量。

## 概率统计

在概率统计中，我们最需要掌握的就是概率的分布。举个例子，我们在做一个预测用户评分的时候，这个分数可能是购买倾向，也可能是信用评分。按照经验，这个评分结果应该是符合正态分布的。这个时候，如果算法同学的模型预测出来的结果不符合正态分布，我们就必须对这个结果进行质疑，让他们给出合理的解释。

从这个例子中我们知道，**概率分布是我们用来评估特征数据和模型结果的武器。** 那产品经理怎么才能利用好这个武器呢？**首先，我们要掌握常用的概率分布的类型。其次，我们还要知道业务场景下的特征数据和模型结果的分布，以及它们应该符合哪种分布类型。这样，产品经理就可以把概率分布应用于日常的工作中。**

因为概率的分布和随机变量的类型相关，随机变量又可以分为离散型随机变量和连续性随机变量两种。为了方便你查看，我把这两种变量对应的概率分布的类型都总结在了下面这张图里。



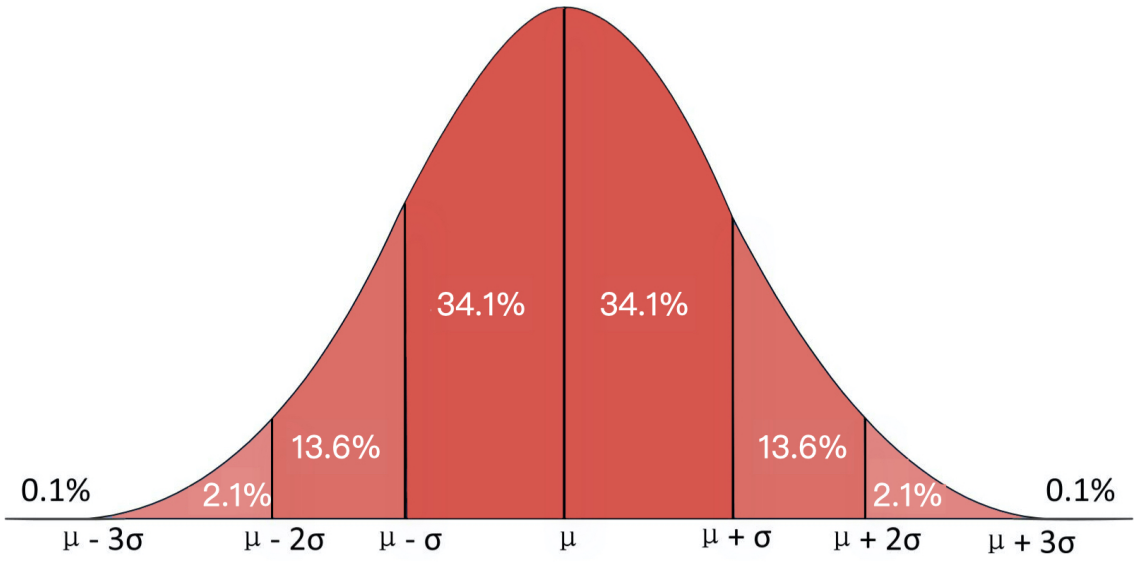

这其中，最常见的概率分布包括伯努利分布、二项分布、泊松分布和正态分布，这也是我们要重点掌握的。下面，我们一一来讲。

**伯努利分布**也可以叫做零一分布。如果我们只进行一次实验，并且这个实验只有两个结果，分别记为 0、1，这就符合伯努利分布。比如在电商场景下，我们设计一个抽奖游戏，某个用户有没有中奖，这个结果就应该符合伯努利分布。

如果重复多次伯努利实验，并且让每次实验都相互独立，让结果只有 0、1 这两种。那 n 次伯努利实验中，结果为 0 的次数的离散概率分布就是**二项分布**。你也可以理解为，抛 n 次硬币，出现正面次数的概率的分布。

**泊松分布描述的是单位时间内，随机事件发生的次数。** 比如，我们的频道页平均每分钟就有 2000 次访问，那如果让我们计算出下一分钟能够有 4000 次访问的概率，这个结果就是泊松分布。

最后，我们来看正态分布，它也叫**高斯分布**。正态分布的曲线特点是**两头低、中间高，左右对称**，所以我们也经常叫它钟形曲线。下图就是一个标准的正态分布图，你看到之后一定会觉得非常熟悉。

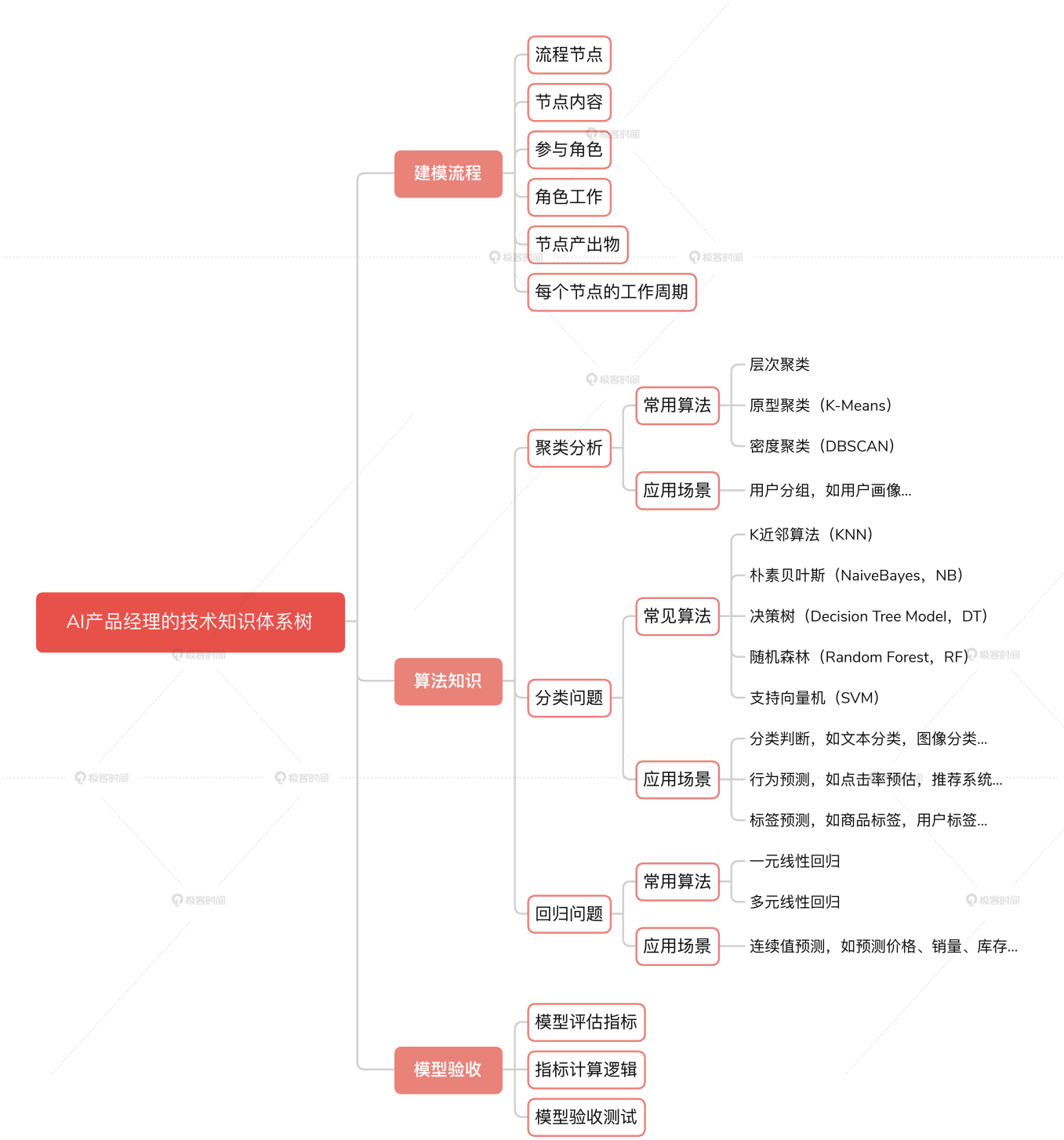



在现实生活中，人的很多特质都符合正态分布，比如人的身高、体重、运动量、智力、收入、甚至信用情况等等。

## 模型构建、算法知识以及模型验收

课程一开始的时候，我就说了，今天对于建模流程、算法知识以及模型验收这三部分内容，只要做到大概的印象就可以，因为这些是后面课程的重点内容。

所以接下来，我会总结出一个知识结构图，给你讲解其中的重要节点，来帮助你形成一个大体的知识框架。这样，当你后面进行学习的时候，再把具体的知识填充到里面，就能形成自己的知识体系树了。




对于模型构建来说，**我们一定要知道模型的建模流程都有哪些节点，这些节点都牵扯了哪些角色，每个角色又承担了什么工作，每个节点的产出物是什么，以及每个节点合理的工作周期又应该是多长时间。**

一个模型构建的整体流程，它一共包括五个阶段，分别为模型设计、特征工程、模型训练、模型验证，以及模型融合。这五个阶段完成之后，模型就会交付到产品端了。

建模的过程实际上就是应用某个算法技术来实现一个模型的过程，这其中最重要的，就是我们选择的是什么算法。所以，作为 AI 产品经理，你要知道目前的技术现状能解决什么问题，在什么场景下有哪些机器学习算法，以及每种算法适合解决哪类问题。

我也按照常用的算法分类方式，把一些常用算法整理在了脑图里。因为算法是未来你区别于普通产品经理的重要抓手，所以我后面会花一整个章节来为你讲解当前主流的算法。

至于模型验收的工作，我也会单独用一个模块来和你详细讲讲。其中，模型验收涉及的评估手段和指标，评估指标背后的计算逻辑，以及怎么选择合适样本进行测试的方法，你都一定要掌握好。

## 模型相关技术名词

在我刚开始转做 AI 产品经理的时候，遇到过这么一件事儿，我们的用户年龄预测模型训练时候 的 KS 值（模型中用于区分预测正负样本分隔程度的评价指标）很高，但是 OOT 测试的时候 KS 还不足 10。当我拿着结果去找算法同学沟通的时候，他就说了一句，“可能是过拟合了，我再改改”。

我刚才提到的 KS、OOT 测试和过拟合你都知道是什么意思吗？如果你知道，说明你的技术基础还不错，如果你还不知道也没关系，当时的我就不知道。

就因为不知道什么是过拟合，所以我什么问题都提不出来，只能自己回去翻书，恶补基础。不过，现在我的算法工程师再和我说出现过拟合问题，我就不会轻易放他离开。因为这明显是模型样本选择出了问题，还有模型验证的时候他也没有认真做，不然这个问题根本不应该到我这里才被发现。他们应该在模型验证的时候就发现这个问题，并且处理掉。

因此，了解一些算法相关的技术名词是非常有必要的，这能帮助我们和算法工程师站在一个频率上对话，提高协作效率和项目效果。

## 过拟合和欠拟合

我们和算法同学聊某个模型表现的时候，他们常常会提到过拟合和欠拟合。它们是什么意思呢？想要理解它们，我们还得先说说偏差和方差是什么。

**偏差**指的是模型的预测结果和实际的结果的偏离程度。如果偏差比较大，就说明模型的拟合程度比较差，也就是**欠拟合**（高偏差），说的直白一些就是模型预测不准。造成欠拟合的原因可能是特征少或者模型训练不足。

而**方差**指的是模型在不同测试样本上表现的稳定程度。假设一个分类模型，在不同样本上测试，得到的 KS 值有时候是 20，有时候是 40，这就说明这个模型方差偏大，模型效果不稳定，在一部分数据上表现好，在另一部分数据上表现差，也就是**过拟合**（高方差）。造成过拟合的原因可能是特征过多或者训练集不够。

## 特征清洗、数据变换、训练集、验证集和测试集

除了欠拟合、过拟合之外，我们还经常听到算法工程师提到当前特征数据质量不好，他们需要花时间进行**特征清洗**。特征数据清洗简单来说，就是对数据进行清洗去掉重复值、干扰数据，以及填充缺失值。一般来说，数据清洗需要反复进行很多次，也会持续很多天，当然具体的工作量也要视数据质量和量级决定。

除此之外，数据有时候还需要进行**数据变换**，处理成方便模型使用的数据形式。举个例子，我们需要使用用户的身高作为模型特征，但是有的数据是用厘米作单位，有的数据会使用米作单位。这个时候，我们就需要使用归一化，把数据的单位统一成米或者厘米。归一化也是数据变换最主要的手段。

最后，关于模型训练我们还需要知道三个概念，**训练集、验证集和测试集**。训练集是让机器学习的样本集合，用来拟合模型。验证集是模型训练过程中，用来对模型性能做初步的评估，用于模型参数调优。测试集是最终用来评估模型效果的。

## 跨时间测试和回溯测试

模型训练结束，我们就要对模型进行测试了。模型测试阶段，我们需要掌握两个关键概念，分别是跨时间测试和回溯测试

**跨时间测试也叫 OOT 测试**，是测量模型在时间上的稳定性。**回溯测试**就是用真实的、过去一段时间的数据，构造出一个模拟的环境（回溯环境），让模型在历史的那段环境中运

行，得到历史某个时间点的模型结果。回溯测试在量化投资中的应用比较广泛。

总的来说，跨时间测试是在模型上线之前就应该要做的事情，回溯测试是指模型已经存在并且已经上线了，我们想要看模型在历史某个时间点的数据表现怎么样的时候，才会进行的测试。

## 联合建模与联邦学习

联合建模和联邦学习的概念你可能还比较陌生，它们经常会在金融领域会用到，而且也是未来建模的一个发展方向。

所谓联合建模，就是使用三方公司（如银联、运营商、电商）的数据，在对方的环境下部署一个模型，然后我们通过接口调用这个模型的结果，再把结果融合到我们自己的模型上。通过 这种方式，可以弥补我们自有业务中数据不足的问题。

但是联合建模会有一个弊端，就是当我们使用三方公司的数据建模之后，在调用的时候，必须传入一个主键来获得模型结果，这个主键如果是用户手机号，身份证号等敏感信息，就会有个人信息泄露的风险。

这个时候，联邦学习就有了用武之地。你可以把联邦学习理解成是特殊的联合建模，或者一种分布式的模型部署方式。使用联邦学习之后，我们调用部署在第三方模型的时候，输入的就不是具体的业务数据而是模型参数，这样就不会有个人信息外传的风险了。目前，蚂蚁、腾讯、京东、微众银行，它们各自都有很成熟的联邦学习解决方案了。

总的来说，如果以后你的业务要使用三方数据，就可以考虑做联合建模。如果想要避免个人信息外传的风险，就可以考虑使用联邦学习技术。

## 总结

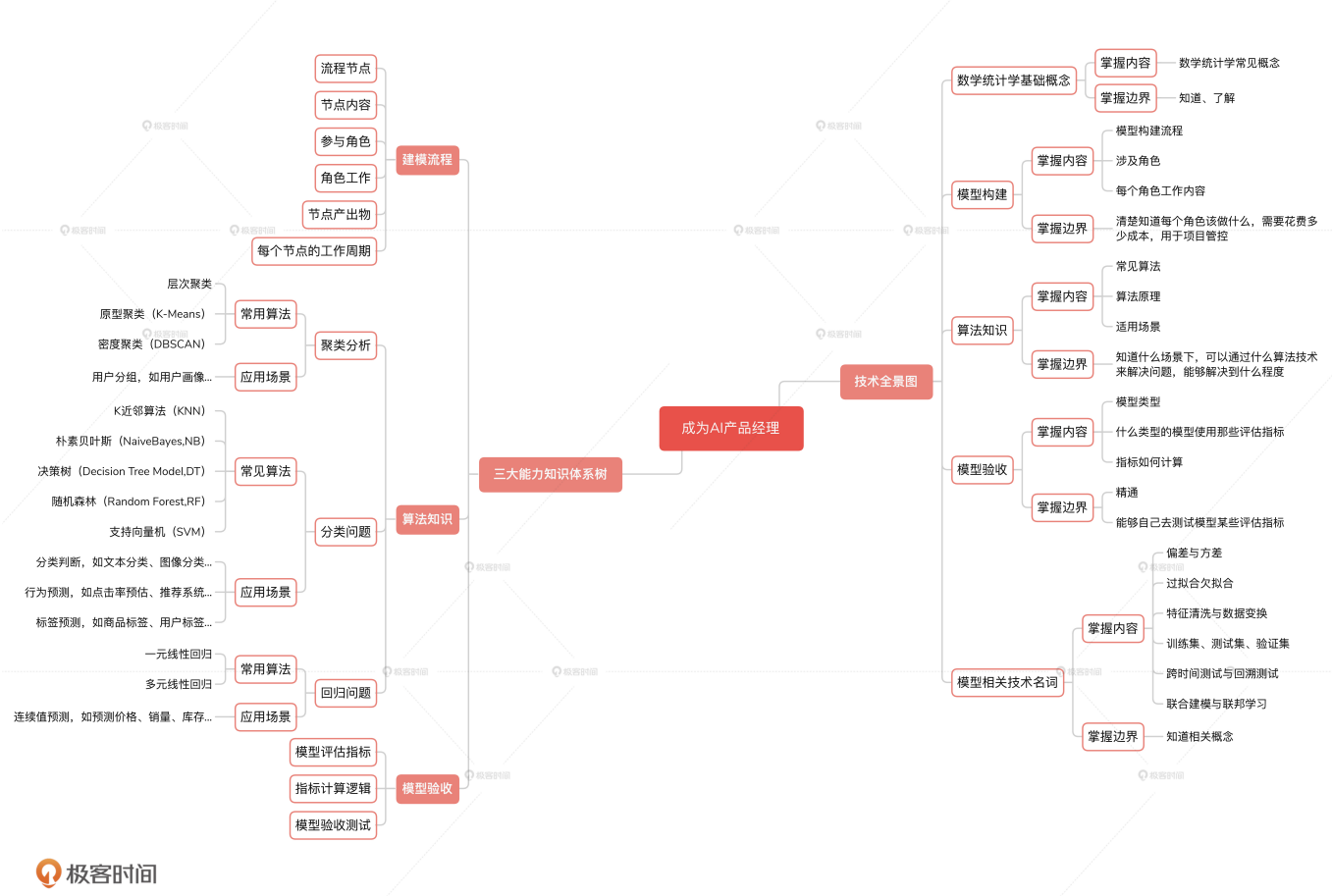
今天，我带你了解了 AI 产品经理应该懂的技术，以及这些技术需要掌握到的程度。

对于数学统计学基础，我们只要掌握今天讲的概念定义就可以； 对于模型构建过程、算法知识和模型验收，你一定要深入了解，知道它们具体的内容和原理；对于模型相关的技术名词，你只要理解我今天列举的常用名词就够了，后面在工作中你可以再慢慢积累，形成你自己的知识体系。

最后，我还想再多说几句，有些同学一看到技术知识，就很容易一头扎进去。但是这些名词在转行初期，你只需要做到知其然就可以。**因为产品经理的职责是善于把技术作为工具和手段，所以我们学习这些专业名词的目的也是为了更好地沟通和推进工作，更好地完成业务目标**，实在没必要把它们的来龙去脉研究透彻。

当然，成功转型后，这些技术知识我们也要进行精进。不过，那个时候，在具体业务的场景下，你就会有更明确的学习方向了。同时，有了算法同学的帮助，你的学习效果也会事半功倍。

我将今天的主要内容总结为了如下所示的思维导图，来帮助你加深理解与记忆。




## 课后讨论

结合你自己的工作经验，你觉得哪些业务场景属于泊松分布，哪些业务场景属于正态分布呢？

期待在留言区看到你的思考和提问，我们下节课见！

提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 02 | 个人视角：成为AI产品经理，要先搞定这两个问题

下一篇 04 | 过来人讲：成为AI产品经理的两条路径

## 精选留言 (2)

写留言

**悠悠** 置顶

2020-12-18

需要什么样的数学基础?

展开

作者回复:
数学基础是很多转型的产品经理都关注的问题。

一个相对完整的数据基础知识包括：线性代数、概率论、数理统计、运筹学（最优化方法）、信息论等等。

首先，产品经理要清晰对于数学基础的掌握边界，与算法和研发相比，产品经理不需要建模，所以有数学基础不需要你究其原理，如数理统计中的统计推断方法，或运筹学中的最优化方法，你的价值在于利用技术服务业务而不是精通技术。

其次，对于想快速入门AI领域的产品经理来说，一定要本着知识的落地应用为学习原则，例如作为产品经理来说为什么要学习概率分布，你要知道它在工作中的作用是什么，其中一点是为了和算法同学同频沟通，另一个很重要原因是为了验收产品。

所以在初期，你大可以以目的为导向只掌握和工作相关的数学知识即可，后面工作稳定下来在适当的去拓展自己的技术边界。

希望能帮到你，悠悠，加油~

2

**Jenny**
2020-12-19

还是不太明白oot，ks是什么？然后为啥oot过高就是模型样本选择的问题呢？求教

作者回复:
oot 是跨时间测试，使用的样本是和模型训练集，验证集没有交集的样本，目的是用来判断模型的泛化能力的。

ks 是判断分类模型的区分能力的，后面会有一节专门来讲什么是 ks，ks 如何计算。

其实不是 oot 过高就是模型样本选择有问题，而是，如果模型在验证集上面 ks 值不错，但是到 oot 测试时候 ks 反而不好，这说明，可能是当时选择的样本不合理，比如：训练和验证样本抽取的时间过早，很多数据发生了变化，导致模型在过去样本上表现好，但是在目前样本上不适用了。

1